8 Analysis of Covariance

Let us recall our previous one-way ANOVA problem, where we compared the mean birth weight (weight) for children in three groups defined by the mother's smoking habits. The three groups had mothers that did not smoke during pregnancy (group 1), mothers that smoked a pack or less of cigarettes per day during their pregnancy (group 2), and mothers that smoked more than one pack of cigarettes per day during their pregnancy (group 3). We concluded that children born to non-smoking mothers were, on average, heavier than children born to mothers in the two smoking groups (and there was no significant difference in birth weights between the two smoking groups).

A deficiency with the analysis is that the differences among groups may be due to other factors that could not be controlled; for example, the mother's intake of caffeine, the mother's pre-pregnancy weight (mweight), and so on. This, of course, is a standard problem with observational studies. If the primary interest is to assess the potential effect of mother's smoking on birth weight, then a proper analysis would need to account for the possible effect of these other features on birth weight. For simplicity, I will consider an analysis that accounts, or adjusts, for the effect of mother's pre-pregnancy weight (mweight) when assessing the effect of smoking. We will see how to adjust for more effects as well.

Let weight_{ij} be the birth weight for the j^{th} child born to a mother in group i ($i = 1, 2, 3$) with pre-pregnancy weight mweight_{ij}. The statistical technique for comparing weights across groups, adjusting for mother's mweight, is called the analysis of covariance (ANCOVA), and is based on the model:

$$\text{weight}_{ij} = \mu + \alpha_i + \beta\,\text{mweight}_{ij} + \epsilon_{ij},$$

where μ is a "grand mean", α_i is the i^{th} group effect, and β is a regression effect. If $\beta = 0$ this is the standard one-way ANOVA model for comparing weights across smoking groups. In words:

$$\text{weight} = \text{Grand Mean} + \text{Group Effect} + \text{mweight Effect} + \text{Residual}.$$

The ANCOVA model implies that the relationship between the mean weight and mother's mweight is linear in each group, but that the regression lines for the groups have different intercepts (and equal slopes). The intercept for group i is $\mu + \alpha_i$. Figure 1 illustrates one possible realization of the model (PPW is mweight).

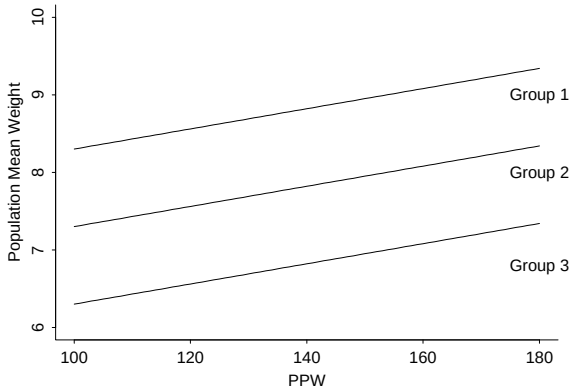

Figure 1: Possible population regression lines for ANCOVA model

Primary interest is in testing the hypothesis of no group effects, which is equivalent to testing that the intercepts of the population regression lines are equal: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. If H_0 is true then the relationship between weight and mweight does not depend on the smoking group to which the mother belongs, that is, there is no effect of mother's smoking on the child's weight, after adjusting for mweight (by including mweight in the model).

Fitting the ANCOVA Model in Stata

ANCOVA is a hybrid of ANOVA and Regression. In **Stata** both the `anova` and `regress` commands assume a continuous response (dependent or y-variable); with `regress` all predictors are continuous, with `anova` all predictors are by default categorical (and a separate indicator variable is created for each level of each predictor). ANCOVA is implemented most easily using the `anova` command (or by using `xi: regress`), but you need to specify what is continuous and what is categorical. The `regress` form is more awkward but is needed when we move to logistic regression. Continuous predictors are known as **covariates**.

Recall the syntax for one-way ANOVA was `anova weight ms_gp` where `weight` is child's birth weight and `ms_gp` is mother's smoking group (Note I coded group as 0,1,2 in the example. This actually serves to illustrate a point, because Stata will in subsequent analysis decide to recode groups as 1-3. I don't want you surprised by this). Page 1 of the **Stata** output has the analysis for this one-way problem. In order to adjust for mweight (Maternal pre-pregnancy weight) the usual method in **Stata** is the command `anova weight ms_gp mweight,cont(mweight)` where the option "`,cont(mweight)`" tells `anova` that `mweight` is continuous. Everything not listed as continuous is assumed to be categorical. You may also use the syntax `anova weight Smoke mweight,cat(Smoke)` to list the categorical variables, with remaining variables treated as continuous.

Page 2 of the **Stata** output begins the summaries of the ANCOVA model. At this time, I will not worry about whether the model fits the data. The F-test for the model gives a p-value for testing no significant effects in the model. The p-value of 0.0000 strongly suggests that either the smoking groups or mweight, or both, have an effect on the weight. The p-values for smoking group and mweight also are both 0.0000, indicating that the group and mweight effects are significant. In particular, there are significant differences in the intercepts of the population regression lines, or put another way, there are significant differences in the mean weights of the three groups defined by mother's smoking habits, **after adjusting for the effect of mweight.** Figure 2 shows the predicted values for the groups from this analysis.

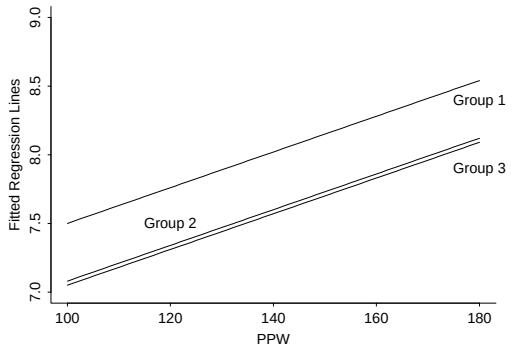

Figure 2: Fitted regression lines for ANCOVA model

A Little More Explanation of the Model

To better understand why ANCOVA is preferred to the one-way ANOVA on birth weights, suppose for argument's sake that weight is strongly positively related to mweight. If smoking behavior is strongly related to mother's mweight, then differences in the mean weights for the three groups could be due solely to differences in mother's mweight. For example, consider the hypothetical population in Figure 3 where I have plotted the relationship between the mean weight and mother's mweight in each group. Suppose that any data collected from these populations falls exactly on the lines in the plot.

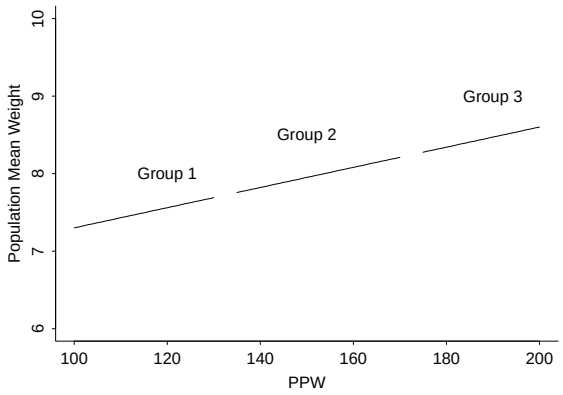

Figure 3: More possible population regression lines for ANCOVA model

A simple linear regression model relating weight to mweight, with no group effects, is appropriate, yet the distributions for weights, ignoring groups, would differ dramatically. The regression line suggests that you would have identical mean weights for mothers in the different smoking groups, if mothers with a fixed weight could be compared across groups.

A one-way ANOVA comparing smoking groups, ignoring mother's mweight would find significant differences in weight across groups. The ANCOVA would conclude, appropriately, that there are no differences across groups once mweight was taken into account. (A more fundamental question is whether these groups, as drawn, are even comparable!)

Interpreting Parameter Estimates

We have already tackled most of the confusing issues with parameter estimates in the two-way ANOVA problem. We have the same issues of constraints here (because we use more model parameters than we actually have "real" parameters). We have 3 intercepts and one slope; we have 1 μ, 3 α_i s, and one β in the model. One of the α_i s is redundant, and **Stata** will take care of that by setting the last one to 0. Again, that is just one possible solution, and different software packages can make different decisions.

We have three lines, and we want the equations for all three. We have forced parallel lines, so the slope is the same for each smoking group, .0118683. The intercept for the i^{th} smoking group has been modelled as $\mu + \alpha_i$, but α_3 has been constrained to 0, so

$$\text{Intercept group } 0 \quad = \quad \text{Constant for model} + \text{gp[0] effect}$$

$$\begin{aligned} &= 5.7517 + 0.4477 = 6.1994 \\ \text{Intercept group 1} &= \text{Constant for model} + \text{gp[1] effect} \\ &= 5.7517 + .0318 = 5.7835 \\ \text{Intercept group 2} &= \text{Constant for model} = 5.7517 \end{aligned}$$

The fitted relationships are

$$\begin{aligned} \text{Predicted weight} &= 6.1994 + .00187 \text{ mweight, for group 0} \\ &= 5.7835 + .00187 \text{ mweight, for group 1} \\ &= 5.7517 + .00187 \text{ mweight, for group 2} \end{aligned}$$

A plot of the fitted relationships is given in Figure 2.

Group Differences

Group differences in the ANCOVA model are differences in intercepts, i.e. vertical distances between the lines. Using a Bonferroni criterion, we see significant differences between Groups 0 and 2 (from `_b[ms_gp[1]]`), and Groups 0 and 1 (from the `lincom (_b[ms_gp[1]]-_b[Smoke[2]])` command), but no significant difference between Groups 1 and 2 (from `_b[Smoke[2]]`). (Note that Stata has renumbered groups as 1, 2, 3 even though we had made them 0, 1, 2 — I apologize for making this confusing, but this "feature" could catch you off guard someday.) This is very similar to what we found in the one-way problem both in direction and size of all effects. Adjusting for mweight has not given us all that much insight here, but needed to be done in order to believe the group differences were "real". The slope of the line has the same interpretation as in regression (what *is* the interpretation?), but is not of primary interest here.

Checking the ANCOVA Model

I have explained the basics of ANCOVA, without considering whether this model describes the CHDS data. The ANCOVA model constrains the slopes of the regression lines for the 3 groups to be identical, so we should check if this is sensible. We can fit three completely different lines (to see if we do any better than by forcing them to be parallel) by adding a smoking group by mweight interaction (crossed effect) to the ANCOVA model.

Pages 4-5 of the **Stata** output show results, as well as three separate linear regressions for the three groups. Let us spend some time in class making sure we see what the parameter constraints have done for us, and that the resulting lines are identical to separate linear regressions.

The p-value of .2491 for the test of no Smoke group * mweight interaction indicates there is no significant improvement by allowing different slopes, so the original ANCOVA model looks reasonable in that regard. The plot of all three fitted lines does not suggest much difference in slopes. The residual plot was not very suggestive of problems – there appears to be little difficulty with this model.

You should note that there is a huge disadvantage to needing a Group*covariate interaction term. With parallel lines we know what group differences mean – the distances between lines. With non-parallel lines the distance between lines depends upon what value of the covariate you are considering, and that distance thus *can be anything you want it to be.* Confirm this with a plot.

Extending the Analysis of Covariance

In the CHDS study, there are several possible effects in addition to mother's pre-pregnancy weight (mweight) that we may wish to consider when assessing whether mother's smoking impacts a child's

birth weight. For example, the mother's height and age, and the gestation length, may be important features to account for in the analysis.

The natural way to account for each effect is through a multiple regression model with a group effect:

$$\text{weight}_{ij} = \mu + \alpha_i + \beta_1 \text{mweight}_{ij} + \beta_2 \text{AGE}_{ij} + \beta_3 \text{HT}_{ij} + \beta_4 \text{GL}_{ij} + \epsilon_{ij}.$$

As before, weight_{ij} is the birth weight for the j^{th} child born to a mother in group i ($i = 1, 2, 3$) with pre-pregnancy weight mweight_{ij}, age AGE_{ij}, height HT_{ij}, and gestation length GL_{ij}.

The multiple regression model has 4 four predictors (mweight, AGE, HT, GL) and 1 factor (groups). The model assumes that the effect of each predictor is the same across groups, leading to a multiple regression model with identical regression effects for each predictor in each group. In words:

$$\text{weight} = \text{Grand Mean} + \text{Group Effect} + \text{mweight Effect} + \text{Age Effect}$$

$$+\text{HT Effect} + \text{GL Effect} + \text{Residual}.$$

A primary interest is testing the hypothesis of no group effects: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. If H_0 is true then the relationship between weight and mweight, AGE, HT, and GL does not depend on the smoking group to which the mother belongs, that is, there is no effect of mother's smoking on the child's weight, after adjusting for mweight, AGE, HT, and GL (by including them as predictors in the model). More generally, the model could include other factors (group variables) or predictors.

Stata Implementation

Six variables are needed to fit this model in Stata: child's weight (continuous), smoking group (1-3, categorical), mother's mweight (continuous), mother's age (continuous), mother's height (continuous), and gestation length (continuous). The command
`anova weight Smoke mweight mage mheight gest,cat(ms_gp)` fits the model. Results are on p. 6 of the output.

The F-tests show all effects but age are important in this model. Confirm that intercepts are -5.838, -6.159, and -6.244 for, respectively, groups 0, 1, and 2. What is of interest of course is the actual group differences. Notice that they are smaller than when ignoring covariates altogether (group differences are the `_b[ms_gp[i]]` terms as before).

Further Thoughts

The approach that I have taken here is consistent with the way epidemiologists assess the impact of a risk factor on a response, adjusting for the effects of confounders. In our analysis, the response is child's weight, mother's smoking habits play the role of a risk factor, and the other features play the roles of confounders (even if they are not strictly so).

A sensible further step in the analysis would be to eliminate, one at a time, the unimportant predictors of weight (i.e. backward elimination). This is easily automated. Once AGE is omitted, the remaining effects are significant at the 1% level. Furthermore, the differences for the smoking groups are nearly identical to those obtained with the previous model, so omitting AGE has little impact on our conclusions. Another reasonable question to examine is whether the smoking groups interact with any of the predictors in the model.

I will note that epidemiologists often adjust for all confounders (at least all that they have measured), regardless of their statistical significance.

Two Simulated Examples

Just to see how much difference covariates can make, I simulated two extreme examples. In the first, the group effect is not at all significant using a simple one-way ANOVA, yet introducing a covariate makes the group effect extremely significant. In this case the covariate is crucial for finding group differences.

In the second example, group differences seem clear from a one-way ANOVA, yet disappear completely when the covariate is introduced. Differences in covariate values completely explain apparent group differences.

We will discuss these examples in class.

Adjusted means

It is very common to report adjusted means in ANCOVA problems. Let us consider the second simulated data set. The table of (raw) means is obtained as follows:

```
. tabstat y,by(group) stat(mean semean)
 Summary for variables: y
     by categories of: group
   group |      mean  se(mean)
---------+--------------------
       1 |  4.472125  .1028853
       2 |   5.50423  .1216309
       3 |  6.472919  .0959465
---------+--------------------
   Total |  5.483091  .1203481
------------------------------
```

We can obtain a table of means adjusted for the covariate x as follows:

```
. adjust x, by(group) se
------------------------------------------------------------
     Dependent variable: y     Command: anova
  Covariate set to mean: x = 3.4761906
------------------------------------------------------------
---------------------------------
    group |         xb        stdp
----------+----------------------
        1 |    5.04042   (.227786)
        2 |    5.50423   (.101776)
        3 |    5.90462   (.227785)
---------------------------------
     Key:  xb    =  Linear Prediction
           stdp  =  Standard Error
```

What is the `adjust` command doing? Match its results with the following:

```
. tabstat x
    variable |      mean
-------------+----------
           x |   3.47619
-------------------------

. lincom(_b[_cons] + _b[group[1]]+_b[x]*3.47619)
 ( 1)  _cons + group[1] + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.040422   .2277854    22.13   0.000     4.584625     5.49622
------------------------------------------------------------------------------

. lincom(_b[_cons] + _b[group[2]]+_b[x]*3.47619)
 ( 1)  _cons + group[2] + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         (1) |    5.50423   .1017763    54.08   0.000     5.300576    5.707883
------------------------------------------------------------------------------
. lincom(_b[_cons] + _b[x]*3.47619)
 ( 1)  _cons + 3.47619 x = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.904621   .2277856    25.92   0.000     5.448823    6.360419
------------------------------------------------------------------------------
```

The problem is that when groups differ on covariate values, the differences in raw means may possibly be attributed just to that, not to real group differences. How can we devise a single summary of a group that is as simple as a mean yet does not suffer this limitation? We fix a single value of the covariate and take the point on the predicted least squares line at that value for each group. This is the estimated mean of the group for that value of the covariate. The default is to take the mean covariate value, but Stata will allow you to take any other value. In this case we are estimating the mean of each population at a value of $x = 3.47619$

Note that differences in adjusted means are just differences in intercepts, and those are the group differences we have been calculating. If we fit an interaction term it is no longer so simple, however. We still can calculate adjusted means, but differences are very specific to the x-value fit.

SAS does the same thing but uses the term LSMEANS (Least Squares Means) instead of adjusted means. The marginal means calculated at the end of the last chapter of notes are more easily obtained using SAS's approach, but can be calculated as adjusted means in Stata as well.

Returning to the preceding example, we see that the adjusted means are not nearly as different as are the raw means. The differences in covariate values explain the apparent differences in groups. The analysis in the separate output is the proper approach, but the table of adjusted means is easier for quick group comparisons on a familiar scale (assuming y has a familiar scale).

Now consider the first simulated data set:

```
. tabstat y,by(group) stat(mean semean)
Summary for variables: y
     by categories of: group
   group |      mean  se(mean)
---------+--------------------
       1 |  4.398373  .2309875
       2 |   4.62633  .2657725
       3 |  4.417013  .2081521
---------+--------------------
   Total |  4.480572  .1347717
----------------------------

. adjust x, by(group) se
-------------------------------------------------
     Dependent variable: y     Command: anova
  Covariate set to mean: x = 3.5
-------------------------------------------------
-----------------------------------
    group |         xb        stdp
----------+------------------------
        1 |    5.41749   (.135878)
        2 |    4.62633   (.114672)
        3 |     3.3979   (.135878)
-----------------------------------
     Key:  xb    =  Linear Prediction
           stdp  =  Standard Error
```

The very small differences in raw means become very large (and have much smaller standard errors when adjusted by x.

You should compute adjusted means in the CHDS data set as an exercise.